Semantic Annotation and Inter-Annotation Agreement in Cast3LB Corpus

Borja Navarro, Raquel Marcos and Patricia Abad

University of Alicante
NLP Research Group (gPLSI)
Departament of Software and Computing Systems
E-mail: borja,rmarcos@dlsi.ua.es,
patriciaabadgarcia@hotmail.com

1 Introduction

In the last few years, new semantically annotated corpora have been developed. Nowadays there are two general topic of interest in semantic annotation of corpora for Natural Language Processing: the first one is the annotation of the unambiguous sense of polysemic words (for example, SemCor [9]). Most of these corpora have been developed for Senseval forum[1] and Word Sense Disambiguation. The second general topic of interest in semantic annotation is the annotation of semantic roles and argumental structures (for example, PropBank [13]).

In this paper we present a new corpus in Spanish annotated with word sense: Cast3LB corpus. It is an “all words” corpus in which all nouns, verbs and adjectives have been annotated with their unambiguous sense.

In the next sections we will present the representation of the semantic information in the corpus and the annotation process. Then we will present the evaluation of the annotation: we have calculated inter-annotation agreement in three experiments. The first one was developed at the beginning of the annotation process, and the objective was to know the level of agreement between annotators without the annotation guide and without any training process. These results have been compared with the results obtained in the second experiment, developed at the end of the annotation process. The comparison shows that agreement increase notably. In this second experiment we have calculated inter-annotation agreement in 13 ambiguous

[1] http://www.senseval.org/

words. The objectives was to know the minimum level of agreement between annotators, that is, the agreement obtained in the annotation of words with high level of complexity. Finally, we have calculated the general agreement between annotator comparing a compleat fragment of the corpus in the third experiment.

Comparing the results obtained with other corpora annotated with word sense, Cast3LB has an inter-annotation agreement similar to the agreement obtained in these other corpora.

2 Cast3LB corpus: annotation overview

Cast3LB corpus is a part of the general multilingual corpus 3LB[2] [15]. This general corpus is formed by three corpora annotated with linguistic information: one for Catalan (Cat3LB), one for Basque (Eus3LB) and one for Spanish (Cast3LB). These three corpora have been annotated at three linguistics levels: morpho-syntactic level, semantic level and discourse-pragmatic level (anaphora).

This paper is focused on the semantic annotation of Spanish corpus Cast3LB. However, these three corpora share the same theoretical framework and the same objectives in semantic annotation.

Cast3LB corpus is a collection of documents extracted from the CLIC-TALP corpus, which is made up of 100.000 words from the LexEsp corpus [16] plus 25.000 words coming from the EFE Spanish Corpus, given by the Agencia EFE (the official news agency) for research purposes. The EFE corpus fragments are comparable among the languages of the general project (Catalan, Basque and Spanish)[3].

We have selected this corpus because it contains a large variety of Spanish texts (newspapers, novels, scientific papers…), both from Spain and South-America, so it is a good representation of the current state of the Spanish language.

As we said before, the corpus is annotated at three linguistic levels:

At morphological level, this corpus was automatically annotated and manually checked in previous projects [4]. At syntactic level, the Spanish corpus and the Catalan corpus have been annotated following the constituency annotation scheme [5]. The bask corpus has been annotated following a dependency annotation scheme [15].

At the semantic level we annotate the sense of nouns, verbs and adjectives, following an “all words” approach. The specific sense (or senses) of each one is assigned by means of the EuroWordNet offset number [23].

[2] Project partially funded by Spanish Government: FIT-150-500-2002-244 and TIC2003-07158-C04-01.

[3] This comparable corpus was developed at Hermes project (http://terral.lsi.uned.es/hermes/)

At the discourse level, we mark the coreference of nominal phrases and some elliptical elements. The coreference expressions taken into account are personal pronouns, clitics, elliptical subjects and some elliptical adjectives. The definite descriptions are not marked. The possible antecedents considered are the nominal phrases or other coreferential expressions [15].

3 Semantic annotation.

As we said before, from a semantic point of view Cast3LB is an "all words" corpus, in which all nouns, verbs and adjectives have been annotated with its specific sense (or senses) in the context in which appear. There has been annotated 42291 words: 20461 nouns, 13471 verbs and 8543 adjectives.

3.1 Representation of semantic information.

For the formal representation of semantic data, we have developed a XML DTD in which, together with the syntactic information, a tag specifies the sense of the words. This specific sense (or senses) of each word is made by means of the EuroWordNet offset number [23], that is, the identification number of the sense (synset) in the InterLingua Index of EuroWordNet.

We have decided to represent the sense of the words by WordNet for several reasons.

- First of all, WordNet is, up to now, the more commonly used lexical resource in Word Sense Disambiguation tasks. It has been used in other semantically annotated corpora like SemCor [9], DSO corpus [11], English "All Words" corpus of SensEval-3 [18] or Italian "All Words" corpus [19]. Indeed, it is the main lexical resource used in Word Sense Disambiguation (see, for example, [10]).

- Secondly, Spanish WordNet is one of the most complete digital lexical resources currently available for Spanish.

- Third, as part of EuroWordNet, the lexical structure of Spanish and the lexical structure of Catalan and Basque are related. Therefore, the annotated senses of the three corpora of 3LB project are also related.

- Moreover, WordNet has a hierarchical semantic structure. This structure allows us to infer semantic generalizations and to achieve more abstract sense than the one represented in the corpus (for example, words related with the same hyponymy).

3.2 Annotation method

The corpus has been annotated by three annotators, but only in some fragments we have followed a double annotation.

It is possible to distinguish two methods for the semantic annotation of corpus based on word sense. The first one is linear (or "textual") method [8], where the human annotator marks the sentences token by token up to the end of the corpus. In this strategy, the annotator must read and analyze the sense of each word every time it appears in the corpus. The second annotation method is transversal (or "lexical") [8], where he/she annotates word-type by word-type, all the occurrences of each word in the corpus one by one. With this method, the annotator must read and analyze all the senses of a word only once.

We have followed in Cast3LB the transversal process ("lexical" method), in which all the occurrences of each word are annotated at the same time by the same annotator. The main advantage of this method is that we can focus our attention on the sense structure of one word and deal with its specific semantic problems: its main sense or senses, its specific senses.... Then we check the context of the single word each time it appears in the corpus and select the corresponding sense. Through this approach, semantic features of each word is taken into consideration only once, and the whole corpus achieves greater consistency.

Through the linear process ("textual" method), however, the annotator must remember the sense structure of each word and their specific problems each time the word appears in the corpus, making the annotation process much more complex, and increasing the possibilities of low consistency and disagreement between the annotators.

Nevertheless, the transversal or lexical method finds its disadvantage in the annotation of large corpus, because no fragment of the corpus is available until the whole corpus is completed. To avoid this, we have selected a fragment of the whole corpus and annotated it by means of the linear process.

Everybody agrees that semantic annotation is a tedious and difficult task. From a general point of view, the main problems in the semantic annotation based on WordNet senses are:

- The subjectivity of the human annotator when it comes to the selection of the correct sense: there are usually more than one sense for a word.

- Due to the WordNet's granularity [14], more than one sense could be correct for a given word.

- The poor agreement between different annotators, due to the ambiguity and/or vagueness of many words.

- Not all nouns, verbs and adjectives are contained in Spanish WordNet. Possible lacks are (i) the synset, (ii) the word, (iii) the synset and the word, and (iv) the link between the synset and the word[4].

In order to overcome these problems, the annotation process has been carried out in two steps. In the first step, a subset of ambiguous words have been annotated twice by two annotators.

With this double annotation we have developed a disagreement typology and an annotation guide. In this annotation guide all the possible causes of ambiguity have been described and common solutions have been adopted for the rest of cases. We try to annotate only one sense per word. If more than one sense could be correct, we annotate the most general sense. The idea is that a general sense includes specific senses. However, there are some cases in which it is necessary to annotate two senses.

In the second step the remaining corpus is annotated following these criteria adopted in the annotation guide.

We have followed a semiautomatic annotation process. In this process, the human annotator must deal only with the polysemic words, because all the monosemic words were annotated automatically[5]. To help the annotation process, a Semantic Annotation Tool (3LB-SAT) has been developed [1].

4 Evaluation: Inter-annotation agreement.

Cast3LB corpus has been annotated semantically by three annotators. Each annotator has annotated a set of words. As we have explained before, due to each word has been annotated completely by the same annotator, the semantic annotation has achieved high consistency.

Only several fragments have been annotated in parallel in order to calculate the agreement between annotators and to know the quality of the semantic annotation.

To do this, we have developed three test or experiments. The first one was a preliminar test. Our objective was to know the difficulty of the semantic annotation task at the beginning of the annotation process. The second experiment was developed at the end of the annotation process. The objective of this experiment

[4] In order to deal with these cases we have defined two more tags: one to express that the word is found in WordNet, but not its correct sense (due to a sense lack, or because there is no link between the word and the synset); and other one to express that the word is not found in WordNet (because it is not there, or because both the word and the synset are missing).

[5] The monosemic words have been checked in order to detect wrong senses.

Word	**occurrences**	**Average**	**Kappa**
Concrete noun: “hombre” (*man*)	12	75%	k = .519
Abstract noun: “vida” (*live*)	13	46%	k = .319
Verb: “decir” (*to say*)	20	25%	k = 0
Adjective: “primero” (*first*)	10	10%	k = 0

Table 1: Inter-annotation agreement. Experiment 1.

was to calculate the minimum level of agreement in words with high level of ambiguity. Finally, the objective of the third experiment was to calculate the general agreement between annotators in a complete fragment of the corpus.

4.1 Experiment 1: preliminar evaluation.

The first experiment was done at the beginning of annotation process. The objective was to know the level of agreement between annotators without the annotation guide and without any training process. The results obtained in this experiment will be compared with the results obtained in the experiment 2.

In this first experiment we have compared the annotation of four specific words with high frequency in the corpus: one verb (“decir”: *to say*), one adjective (“primero”: *first*), one abstract noun (“vida”: *life*) and one concrete noun (“hombre”: *man*).

All occurrences of these words in the corpus were annotated by two annotators. The results are shown in Table 1.

The results show that the agreement between annotators at the beginning of the annotation process is very low. Indeed, in the annotation of the verb and the adjective, the agreement according to kappa coefficient (see later) is 0, that is, the agreement between annotators is the agreement expected by chance.

On other hand, as is shown in [21], the level of ambiguity is different in each part of speech. Effectively, according to this preliminar experiment, adjectives are more ambiguous than verbs and nouns; and abstract nouns are more ambiguous than concrete nouns.

4.2 Experiment 2: minimum level of agreement.

The second experiment was developed at the end of the annotation process, once the annotation guide was developed and annotators have experience in semantic annotation.

The objective of this second experiment was to know the minimum level of agreement between annotators; that is, to know the agreement achieved in complex words: words with high level of ambiguity.

Word	**Part of Speech**	**Senses**	**Frequency**
Historia (*history*)	Noun	9	33
Carrera (*race*)	Noun	11	27
Ley (*law*)	Noun	6	22
Tierra (*earth*)	Noun	11	18
Papel (*paper*)	Noun	7	18
Ganar (*to win*)	Verb	8	33
Suponer (*to suppose*)	Verb	10	33
Pensar (*to think*)	Verb	8	38
Trabajar (*to work*)	Verb	8	33
Jugar (*to play*)	Verb	7	26
Nacional (*national*)	Adjective	10	26
Abierto (*open*)	Adjective	28	17
Personal (*personal*)	Adjective	10	20

Table 2: Words and frequency

Word	**Average**	**Agreement by chance (P(e))**	**Kappa**
Historia (*history*)	45%	0.23	k = .28
Carrera (*race*)	89%	0.43	k = .8
Ley (*law*)	75%	0.266	k = .66
Tierra (*earth*)	56%	0.17	k = .46
Papel (*paper*)	78%	0.42	k = .61
AVERAGE	68%	-	k = .56

Table 3: Minimum inter-annotation agreement in nouns. Experiment 2.

The inter-annotation agreement has been calculated following an evaluation method of “lexical sample” corpus. The annotation of 13 ambiguous words (5 nouns, 5 verbs and 3 adjectives -see table 2-) has been compared. The results are shown in table 3, 4 and 5.

The average of agreement between the three part of speech are 68%. Similar to the first experiment, the part of speech with less agreement is the adjective (63%). However, verbs are the part of speech with higher level of agreement (72%).

Together with the agreement average, we have calculate the kappa agreement, following [17] [6]. This measure calculates and removes from the agreement rate the amount of agreement that is expected by chance. Therefore, the result are more exact than a simple agreement average [2] [22] [6].

Kappa measure is calculated according to the formula:

[6] There are two method for calculate kappa [7]: one in [6] and the other one in [17]. We have used both formula. The results are quite similar, so we show here only the results obtained with Siegel and Castellan formula [17].

Word	**Average**	**Agreement by chance (P(e))**	**Kappa**
Ganar (*to win*)	87%	0.66	k = .61
Suponer (*to suppose*)	28%	0.25	k = .15
Pensar (*to think*)	89%	0.45	k = .8
Trabajar (*to work*)	71%	0.54	k = .36
Jugar (*to play*)	76%	0.3	k = .65
AVERAGE	72%	-	k = .51

Table 4: Minimum inter-annotation agreement in verbs. Experiment 2.

Word	**Average**	**Agreement by chance (P(e))**	**Kappa**
Nacional (*national*)	62%	0.45	k = .3
Abierto (*open*)	50%	0.14	k = .41
Personal (*personal*)	41%	0.31	k = .15
AVERAGE	63%	-	k = .29

Table 5: Minimum inter-annotation agreement in adjectives. Experiment 2.

$$k = \frac{P_A - P_E}{1 - P_E}$$

where P_A represents the annotation agreement percentage and P_E the agreement by chance.

The kappa measure obtained is K =.45. To obtain this result, we have calculated the kappa measure of each word independently, then we have calculated the average between all words of the same part of speech, and, at the end, we have calculated the average between the three part of speech. In [3], this average is called "macro-kappa".

Comparing these results with the results obtained in the first experiment, the inter-annotation agreement increases notably with the training process and with the annotation guide.

The results are similar to the results reported in [21] and other corpora as [3], [12]. Table 6 shows a comparison between Cast3LB corpus and these corpora that have used kappa to measure inter-annotation agreement. This table shows the agreement achieved only in nouns.

It must be noted that the amount of words used in Cast3LB to calculate kappa agreement is less than the amount of words used in the other corpora. For example, [3] use 280 words to calculate kappa. We have used only 13 words. Maybe, this is the reason why Cast3LB shows high level of agreement. In any case, the results show that inter-annotation agreement in Cast3LB is quite similar to other corpora.

Corpus	kappa
Chklovski & Mihalcea [3]	.35
Ng [12]	.30
Véronis [20]	.49
Cast3LB	.56

Table 6: Comparing kappa agreement in several corpora.

Part of Speech	Amount of words	Total agreement	Average
Nouns	327	254	77.67%
Verbs	147	103	70%
Adjectives	73	59	80.82%
TOTAL	547	416	76.05%

Table 7: General inter-annotation agreement. Experiment 3.

4.3 Experiment 3: general agreement.

In the third experiment the inter-annotation agreement was calculated again. However, in this case we want to calculate the inter-annotation agreement with a complete passage of the corpus, not only for complex words. As we said before, Cast3LB is an "all words" corpus, in which all nouns, verbs and adjectives have been annotated. It is different from "lexical sample" corpus, in which only some ambiguous words are annotated.

In this third experiment we have calculated the inter-annotation agreement as an "all word" corpus, that is, we have calculated the inter-annotation agreement for all nouns, verbs and adjectives in a complete passage of the corpus. Unlike second experiment, in this third experiment we have a lot of different words, but with few occurrences each one.

We have compared two files annotated in parallel by two annotators. In these files there are 548 words: 74 adjectives, 327 nouns and 147 verbs. There are total agreement in 516 words: 60 adjectives, 256 nouns and 103 verbs. The average agreement for adjectives was 81.08%, for nouns 78.28% and for verbs 70.06% (Table 7).

According to this results, the general agreement in Cast3LB is 76%.

These results are similar to the agreement obtained in others "all words" corpora. Table 8 shows a comparison between some of the main "all words" corpora. Cast3LB has a level of agreement similar to these other corpora[7].

In this third experiment we have not applied kappa measure to calculate inter-

[7] As we said before, it must be noted that the amount of words used in Cast3LB to calculate general agreement is less than the amount of words used in the other corpora.

Corpus	**General average**
DSO [11]	80 - 90% aprox.
Cast3LB	76.05%
SEMCOR [9]	73%
English “all words” in Senseval-3 [18]	72.5%
Chklovski & Mihalcea [3]	67.3%

Table 8: Inter-annotation agreement in several “all words” corpora.

annotation agreement. To apply kappa to the evaluation of an “all words” corpus has several problems:

1. Firs of all, the data to be compared in an “all words” corpus are not homogeneous. When kappa measure is applied to inter-annotation agreement in word sense annotation, the possible senses of a word are the classes in which the word could be classified. These classes must be homogeneous. In an “all words” corpus they are not homogeneous because there are words with one sense, words with two, three senses, etc. Only the annotation of word with the same amount of senses could be compared with kappa measure.

2. However, the senses of two ambiguous words with the same number of senses are not comparable: the senses of some words are very different (so it is easy to select between them the correct one) and the senses of other words are very similar (so it is very difficult to annotate these words). This differences in the senses are not reflected in kappa statistics. For kappa, all senses (all classes) are equal.

[3] proposes a solution to these problems. They propose calculate two averages: micro-kappa average and macro-kappa average. The first one is based on the average of all words (micro-kappa) and the second one is based on the average of the kappa measure of each word (macro-kappa).

In order to evaluate the semantic annotation of Cast3LB, in this paper we propose two measures of agreement: minimum level of agreement and general agreement. Minimum level of agreement is calculated based only on words with high level of ambiguity. For this measure kappa is used. General agreement is calculated for all words annotated in a sample of the corpus (nouns, verbs and adjectives).

5 Conclusions

In this paper we have shown the semantic annotation of Cast3LB and the inter-annotation agreement achieved. We have developed three experiments in order to specify a correct evaluation. First, we have developed a preliminar evaluation in order to know the agreement between two annotators without experience and without an annotation guide. In the second experiment, inter-annotation agreement between several ambiguous words has been calculated following a “lexical sample” approach. Due to these words are words with high level of ambiguity, this experiment shows the minimum level of agreement. Finally, in the third experiment we have calculated the general agreement between annotators in a complete passage of the corpus. In this experiment, all words annotated have been compared. The results of all these experiments show a level of agreement similar to other “all words” corpora.

References

[1] Bisbal, E.; Molina, A.; Moreno, L.; Pla, F.; Saiz-Noeda, M. (2003) 3LB-SAT: una herramienta de anotación semántica, In *Procesamiento del Lenguaje Natural*, 31.

[2] Carletta, Jean. (1996) Assessing Agreement on Classification Tasks: The Kappa Statistics. *Computational Linguistics*, 22.

[3] Chklovski, Timothy and Mihalcea, Rada (2003) Exploiting Agreement and Disagreement of Human Annotators for Word Sense Disambiguation, In *Proceedings of Recent Advances In NLP (RANLP 2003)*. Bulgaria.

[4] Civit, M. (2003) *Criterios de etiquetación y desambiguación morfosintáctica de corpus en Español*. Alicante, Sociedad Española para el Procesamiento del Lenguaje Natural.

[5] Civit, Montserrat; Martí, MA Antonia; Navarro, Borja; Bufí, Núria; Fernández, Belén and Marcos, Raquel (2003) Issues in the Syntactic Annotation of Cast3LB *4th International on Workshop on Linguistically Interpreted Corpora (LINC-03), 10th Conference of the European Chapter of the Association of Computational Linguistics (EACL)* Budapest.

[6] Cohen, Jacob. (1960) A coeficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20.

[7] Di Eugenio, Barbara and Glass, Michael. (2004) The Kappa Statistic: A Second Look, In *Computational Linguistics*, 30(1).

[8] Kilgarriff, Adam. (1998) SENSEVAL: An Exercise in Evaluating Word Sense Disambiguation Programs. In: *Proceedings of Language Resource and Evaluation Conference*.

[9] Miller, G. A., Leacock, C., Randee, T., and Bunker, R. (1993) A Semantic Concordance. In: *Proceedings of the 3rd ARPA Workshop on Human Language Technology.*

[10] Mihalcea, R.; Chklovsky, T.; Kilgarriff, A. (2004) The Sensenval-3 English lexical sample task. *Senseval-3. Third International Workshop on the Evaluation os Systems for the Semantic Analysis of Texts. ACL-04*, Barcelona.

[11] Ng, H. T., and Lee, H. B. (1996) Integrating Multiple Knowledge Sources to Disambiguate Word Sense: An examplar-Based Approach. In: *Processding so the Association of Computational Linguistics.*

[12] Ng, H.T., Chung, Y. L. and Shou, K. F. (1999) A Case Study on Inter-Annotation Agreement for WSD. In *Proceedings of the SIGLEX Workshop "Standardizing Lexical Resources"*, Maryland, USA.

[13] Palmer, Martha ; Gildea, Daniel, and Kingsbury, Paul (2005) The Proposition Bank: An Annotated Corpus of Semantic Roles, In: *Computational Linguistics*, Vol. 31 (1), pp. 71 - 105.

[14] Palmer, M. (1998) Are WordNet sense distinctions appropriate for computational lexicons? *Proceedings of Senseval, Siglex98.* Brighton, England.

[15] Palomar, M.; Civit, M.; Díaz, A.; Moreno, L.; Bisbal, E.; Aranzabe, M; Ageno, A.; Martí, M.A. and Navarro, B., (2004) 3LB: contrucción de una base de datos de árboles sintáctco-semántico para el catalán, euskera y castellano, *Procesamiento del Lenguaje Natural*, 33.

[16] Sebastián, N., Martí, M.A., Carreiras, M. F., and Cuetos, F. (2000) *LEXESP: Léxico Informatizado del Español.* Barcelona: Edicions de la Universitat de Barcelona.

[17] Siegel, S. and Catellan, N. J. (1988) *Nonparametric Statistics for the Behavioral Science*, Boston, McGraw-Hill.

[18] Snyder, Benjamin, and Palmer, Martha. (2004) The English All-Word Task. In: *Porceedings of SENSEVAL-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text. ACL-04*, Barcelona.

[19] Uliveri, M.; Guazzini, E.; Bertagna, F,; and Calzolari, N. (2004) Senseval-3: The Italian All-words Task. In: *Proceeding of Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Anlysis of Texts. ACL-04* Bacelona.

[20] Véronis, J. (2000) Sense tagging: Don't look for the meaning but for the use, In *Computational Lexicography and Multimedia Dictionaries (COMLEX'2000),* Kato Achia (Greece), pp. 1-9.

[21] Véronis, Jean (2001) Sense tagging: does it make sense? In *Corpus Linguistics Conference*, Lancaster, U.K.

[22] Vieira, Renata (2002) How to evaluate systems against human judgment on the presence of disagreement?

[23] Vossen, Piek. (2002) *EuroWordNet General Document. Part A. Final Document.* EuroWordNet (LE2-4003, LE4-8328).